cs523 - Big Data Technology

Extra Credit Lab 10 – Apache Spark

Submit your *own work* on time. No credit will be given if the lab is submitted after the due date. Follow the instructions completely.

Analysis on Apache access log file

A sample Apache access log file is given to you.

Get some information about Apache log file format from here and refer to this link for parsing this log file.

Note that these log file formats keep changing and the Regex that is available at the above link might be a little different from the given file. So, figure it out yourself!

Find out the following information from the file by writing Java Spark programs.

1. **[5] Find out how many 401 errors are there within a particular date range. Assume that the date range will be supplied by the user.**
   You can take the date range as runtime argument.

   Submit your java file, output file, date range and the command used to run the program in pseudo-distributed mode.

2. **[5] List all the IP addresses that have accessed this server more than 20 times.**

   Submit your java file, output file and the command used to run the program in pseudo-distributed mode.